Summarization of an Online Medical Encyclopedia

Marcelo Fiszman, Thomas C. Rindflesch, Halil Kilicoglu

National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA

Abstract

We explore a knowledge-rich (abstraction) approach to summarization and apply it to multiple documents from an online medical encyclopedia. A semantic processor functions as the source interpreter and produces a list of predications. A transformation stage then generalizes and condenses this list, ultimately generating a conceptual condensate for a given disorder topic. We provide a preliminary evaluation of the quality of the condensates produced for a sample consisting of four disorders. The overall precision of the disorder conceptual condensates was 87%. The compression ratio from the base list of predications to the final predications in the condensate was 98%. The final conceptual condensate is not an actual summary; however, it could be used as input to a text generator to produce a natural language summary for a given disorder topic.

Keywords

Natural Language Processing, Automatic Summarization, Knowledge Representation.

Introduction

The amount of information available online is growing exponentially. Paradoxically, the more resources grow, the harder it is for users to access information efficiently. Automatic text summarization is an enabling methodology that presents users with compressed yet reliable information.

Spark Jones [1] defines a summary as "a reductive transformation of source text to summary text through content reduction selection and/or generalization on what is important in the source." She further describes the automatic summarization process in three stages:

1. *Interpretation* of the source into source text representation;
2. *Transformation* of source representation into summary text representation and;
3. *Generation* of summary text from transformed summary representation.

The crucial issue in this framework is the information that must be included in the summary text representation in order to create an adequate summary. This is addressed largely in the transformation stage, which attempts to condense the source text representation. Several methodologies and architectures have been proposed for automatic summarization, and they can be broadly separated into extraction and abstraction methods [1, 2].

In this paper, we explore an abstraction methodology and apply it to multiple documents from an online medical encyclopedia. We rely on a natural language processing system (called SemRep) and a transformation stage to produce conceptual condensates for disorder topics. We do not generate a summary text but display the summarized information in graphical format. Finally, we provide a preliminary evaluation of the quality of the condensates produced for four disorders.

Background

Automatic Text Summarization Research

The extraction paradigm focuses on identifying salient sentences, which are determined by assigning weights based on such features as location in text, frequency of occurrence, cue phrases, and statistical relevance measures [3, 4]. The overall salience is computed for each sentence and the best ones are kept as a summary. This approach is sometimes called knowledge-poor, since it does not rely on meaning or language structure.

According to Hahn [2], there are two abstraction approaches, and both are knowledge-rich. The first relies heavily on syntactic parse trees for [5] producing a structural condensate. The second approach also uses natural language processing, but the final source text representation is conceptual rather than syntactic. The transformation phase is a condensation and generalization operation that operates on this semantic conceptual space, eliminating redundant information, merging graphs, establishing connectivity patterns, and enforcing compliance with summary-oriented inference rules [6]. The final representation is a *conceptual condensate* of the original text.

We follow the conceptual abstraction paradigm, which has not been extensively explored because of its heavy dependence on domain knowledge. We rely directly on the Unified Medical Language System® (UMLS) ® [7]. Although the UMLS knowledge sources are not intended as ontologies and will not support extensive inferencing without enhancement, they provide breadth of coverage of the biomedical domain. SemRep is used as the source interpreter and the transformation stage operates on the semantic predications produced by SemRep to summarize information about disorders.

UMLS Resources

All three UMLS knowledge sources, the Metathesaurus,® the Semantic Network, and the SPECIALIST Lexicon are used by SemRep. An interpreter for hypernymic propositions (predications where the arguments are in taxonomic relation) has been recently added to SemRep. It relies heavily on semantic groups from the Semantic Network and hierarchical relationships from the Metathesaurus.

McCray [8] reduces the conceptual complexity of medical knowledge represented in the Semantic Network through the use of semantic groups, which organize the 134 semantic types in the Semantic Network into 15 coarse grained aggregates. As an example the semantic group **Disorders** contains such semantic types as 'Disease or Syndrome', 'Neoplastic Process', and 'Mental or Behavioral Dysfunction'.

Materials and Methods

Our automatic summarization process is illustrated in Figure 1 and described below.

The Source: The Online Medical Encyclopedia

The A.D.A.M. Health Illustrated Encyclopedia® [9], which is available through The National Library of Medicine's MEDLINE*plus*® includes over 4,000 entries on diseases, tests, symptoms, injuries, and surgeries. It also contains an extensive library of medical photographs and illustrations. Each article has a main topic followed by free text information on that topic. Taking disease as an example, a typical entry has information about that particular disease (definition, causes, incidence, risk factors, symptoms, treatment, etc.) The medical photographs and illustrations have captions in free text format in separate Web pages, increasing the number of pages to approximately 5,000.

The Interpretation Stage: SemRep

SemRep [10] identifies semantic propositions in biomedical text, and we used it as the source interpreter for this project. During processing, an underspecified syntactic parser depends on lexical look-up in the SPECIALIST lexicon and the Xerox Part-of-Speech Tagger. MetaMap [11] matches noun phrases to concepts in the Metathesaurus and determines the semantic type for each concept. Argument identification is based on dependency grammar rules that enforce syntactic constraints. Indicator rules map syntactic phenomena to predicates in the Semantic Network, which imposes semantic validation for the associative relationships constructed. As an example, consider (1)

(1) **Proton pump inhibitors** are now the first choice in the *treatment of* **Zollinger-Ellison syndrome**

A semantic indicator rule links the nominalization *treatment* with the Semantic Network predicate "Pharmacologic Substance-TREATS-Disease or Syndrome." Since the semantic types of the syntactic arguments identified for *treatment* in this sentence match the corresponding semantic types in the predication from the Semantic Network, the predication (2) is constructed.

(2) Proton pump inhibitors-TREATS-Zollinger-Ellison syndrome

SemRep has a module [12] that interprets hypernymic propositions. After syntactic arguments have been identified, MetaMap matches them to concepts in the Metathesaurus. Such concepts are then subjected to semantic validation. The semantic types must occur within the same semantic group and the concepts themselves must be in a hierarchical relationship in the Metathesaurus. As an example, consider the instance of a nominal modification highlighted in (3).

(3) The [**antibiotic tetracycline**] given before the age of 8 years can cause abnormal tooth color.

Based on the underspecified parse, in which the head and the modifier are identified for the noun phrase in bold, MetaMap identifies the Metathesaurus concepts "Antibiotics" and "Tetracycline" and their respective semantic types ('Antibiotics' in both cases). Since the semantic types belong to the semantic group Chemicals & Drugs, the Metathesaurus hierarchical file is consulted, and it is determined that "Antibiotics" is an ancestor of "Tetracycline," thus allowing the construction of the logical form in (4).

(4) Tetracycline -ISA- Antibiotics.

We processed the 5,000 Web pages from the encyclopedia with SemRep to produce a set of predications in logical form (subject-PREDICATE-object) at the sentence level. SemRep does not resolve anaphoric expressions at the discourse level. Therefore, no attempt was made to take advantage of rhetorical structure for summarization [13].

Before the transformation stage begins, predications are subjected to a word sense disambiguation filter. From previous work in SemRep, word sense ambiguity was one of the major causes of false positive mistakes. Branded drug

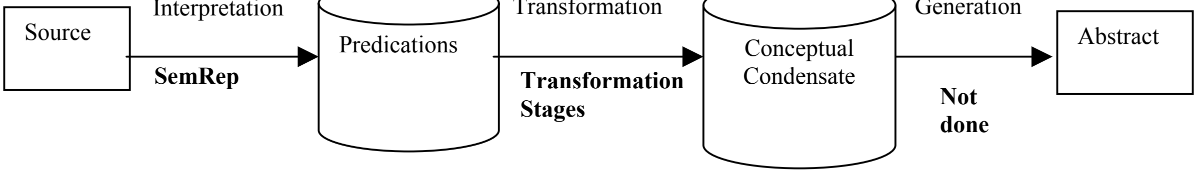

Figure 1 - The summarization paradigm used in our methodology as proposed by Hahn and Jones. The source is a multiple-document source (Online Medical Encyclopedia). The generation stage was not performed in this study.

names such as Duration (Duration brand of oxymetazoline), Direct (Direct type of resin cement), and others, which are ambiguous with the more common sense of their names are a particular problem. 37,281 unique predications were generated while processing the encyclopedia, and the word sense disambiguation filter reduced this list to 36,608.

The Transformation stage

The transformation stage takes as input the list of unique filtered predications and a seed disorder concept, which has to be a Metathesaurus concept. The output is a conceptual condensate for the input disorder concept. In the abstraction paradigm the transformation stage condenses and generalizes [2], and in our knowledge-rich approach, these processes are conducted on the list of predications obtained from SemRep, in three stages.

Stage 1, a condensation process, identifies predications on a given topic (in this study, disorders) guided by a semantic schema for that topic. This provides a set of core predications on that topic. Stage 2 is a generalization process and identifies non-core predications occurring in the neighboring semantic space of the core predications. This is accomplished by retrieving all the predications that share an argument with one of the core predications. Stage 3 provides further condensation by eliminating predications with generic arguments based on hierarchical information from the UMLS.

Stage1 - Disease description schema

We base our schema for disorders on disease description frames as proposed recently by Jacquelinet and Burgun [14]. In adapting their frames to SemRep predications in the form, subject-PREDICATE-object, the following predicates with their respective domains are used:

{Disorders} ISA {Disorders}
{Etiological process} CAUSES {Disorders}
{Treatment} TREATS {Disorders}
{Body location} LOCATION_OF {Disorders}
{Disorders} OCCURS_IN {Disorders}
{Disorders} CO-OCCURS_WITH {Disorders}

{Disorders} is a subset of the semantic group **Disorders** and contains the following semantic types: {'Disease or Syndrome', 'Neoplastic Process', 'Mental or Behavioral Dysfunction', 'Sign or Symptom'}.

{Etiological process} is a subset of the union of two semantic groups **Living Beings** and **Chemicals & Drugs**. It contains the following semantic types: {'Bacterium', 'Virus', 'Fungus', 'Invertebrate', 'Rickettsia or Chlamydia', 'Amino Acid, Peptide, or Protein', 'Biologically Active Substance', 'Element, Ion, or Isotope', 'Hazardous or Poisonous Substance', 'Antibiotic', 'Pharmacologic Substance', 'Immunologic Factor', 'Organophosphorous Compounds'}.

{Treatment} is a subset of the union of two semantic groups **Chemicals & Drugs** and **Procedures.** It contains the following semantic types: {'Pharmacologic Substance', 'Antibiotic', 'Hormone' 'Vitamin', 'Therapeutic or Preventive Procedure'}.

{Body location} is a subset of the semantic group **Anatomy** and contains the following semantic types: {'Body Part, Organ, or Organ Component', 'Body Location or Region', 'Body Space or Junction', 'Fully Formed Anatomical Structure'}.

In addition, we allowed hypernymic predications (ISA) for {Etiological process} ISA {Etiological process} and {Treatment} ISA {Treatment}. We have not done so for {Body location}, because these are meronomic (PART_OF) relations and our interpreter only deals with taxonomic relationships.

Although the predicates, domains, and semantic types allowed for the domains are not complete, they represent a substantial amount of what can be said about disorders.

During processing in the first stage, all predications are retrieved from the input list that have the seed concept as an argument and that conform to the restrictions of the template. This forms a core list of predications about the seed concept. For example, Stage 1 processing with the seed concept "Asthma" retrieves predications such as Asthma-ISA-Obstructive Lung disease, Allergens-CAUSE-Asthma, Asthma-CO-OCCURS_WITH-Bronchiolitis, and Albuterol-TREATS-Asthma.

Stage 2 - Connectivity

The connectivity stage generalizes the conceptual condensate by expanding the core list of predications to neighboring semantic space. It does so by examining all non-seed concepts in core predications and finding additional predications that contain that concept. For example, from the core predication Albuterol-TREATS-Asthma, predications containing the non-seed concept, "Albuterol," such as Albuterol-ISA-Bronchodilator Agent are retrieved. Currently, the system extends the list of predications with this chaining process to only once; it does not recurse on the non-core predications.

Stage 3 - Hierarchical principle

The final stage in the transformation process eliminates uninformative predications having a generic argument such as "Pharmaceutical Preparations" or "Disease." This is accomplished by examining the hierarchical position of each concept in all predications in three medical terminologies: Clinical Terms Version 3 1999 (Read Codes), Computer Retrieval of Information of Scientific Projects 2003 (CRISP Thesaurus), and Medical Subject Headings 2003 (MeSH).

The distance between the concept and the root is calculated for each source. A set of rules was developed for each domain of the disorder schema, and empirically-determined values indicate when to prune in each source. For example, an argument in the {Disorders} domain is pruned if the distance to the root is less than four in Read Codes or less than three in the CRISP Thesaurus.

Evaluation

Evaluation in automatic summarization attempts to measure either the quality of the summary as related to the source or how the summary affects the completion of some other task [15]. We have not yet addressed these issues; however, we have conducted a preliminary evaluation of the quality of the conceptual condensates generated for four disorder concepts (Gout, Hyperthyroidism, Migraine and Chest Pain).

The first author (MF) examined the source sentence that SemRep used to generate each predication in these condensates and marked the predications as either correct or incorrect. Precision was calculated as the total number of correct predications divided by the total number of predications in the condensate.

We also measured the amount of reduction (compression) in the number of predications for each of the four seed concepts. The base number of predications is calculated after the connectivity (generalization) stage is applied. The final number of predications is determined after the final transformation stage (hierarchical) has been applied.

Results

Results for the quality of the conceptual condensates and reduction in the list of predications for each of the four seed concepts are shown in Table 1.

Table 1 - Results for the four seed disease concepts.
***C** = Correct, **I** = Incorrect.*

Concept	Base	Final	C	I	Precision
Chest Pain	1270	31	26	5	83%
Gout	2832	37	33	4	89%
Hyperthyroidism	3224	51	47	4	92%
Migraine	2726	71	60	11	84%
Total	**10052**	**190**	**166**	**24**	**87%**

The compression rate from the base to final number of predications was approximately 98%. Out of 190 total final predications, the distribution of predications for each of the six predicates allowed in the disease templates was: ISA–64, CAUSES-21, TREATS-43, LOCATION_OF-23, OCCURS_IN-14, CO-OCCURS_WITH-35. Figure 2 is the resulting conceptual condensate for the seed chest pain visualized using Pajek, a software for analysis and visualization of large networks [16].

Discussion

The conceptual condensates produced by our summarization processing provide an overview of the seed disorder, including characteristics and treatments. Using Figure 2 as an example for the seed topic, chest pain, one can paraphrase what was extracted from the encyclopedia. Noting the relations between the nodes, we see, for example, that Propranolol, which is an adreneregic beta-antagonist, Verapamil, which is a calcium channel blocker, and nitroglycerin can treat chest pain. The condensate also contains the statement that Verapamil can treat angina pectoris, which is a kind of chest pain. Further, it can be seen that chest pain is a symptom, which co-occurs with insomnia, sweating and palpitations.

False positives were mainly due to word sense ambiguity and incorrect argument identification by SemRep. One example of word sense ambiguity related to the way information is represented in the UMLS can be seen in (5).

(5) **Propranolol** is used for **hypertension**

Hypertension has two senses and one of them is 'Hypertension induced by pregnancy'. The latter one is chosen as the object argument of the TREATS predicate. Word sense disambiguation is still a matter of investigation in natural language processing.

A limitation of our evaluation is that we did not determine recall errors. Since our source consists of multiple documents, it would be difficult to evaluate completeness of the condensate, because assertions can come from any sentence in the source. Another limitation is that only one person evaluated the quality of the condensate. Other evaluation studies in text summarization [15] have shown that inter-rater reliability is an issue. From a methodological perspective, we did not address rhetorical structural analysis, which is important in summarization research [13].

In this study, we concentrated on disorders. We would like to expand this methodology to other topics, and we would like to use a text generator to complete the task and have an actual natural language summary for the seed concept.

Conclusion

We have presented a knowledge-rich (abstraction) approach to summarization of multiple documents from an online medical encyclopedia. Our abstraction approach uses a natural language processing system and a transformation stage targeting disorder topics in order to produce conceptual condensates for disorders. The compression rate for the predications was high and the quality of the condensates was good. Precision in this sample was 87%.

Acknowledgments

The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

[1] Jones KS. Automatic Summarizing: Factors and Directions. In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999; pp 1-13.

[2] Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29-36.

[3] Hovy E, Lin CY. Automated Text Summarization in SUMMARIST. In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999; pp 81-94.

[4] Teufel S, Moens M. Summarizing Scientific Articles - Experiments with Relevance and Rhetorical Status. Computational Linguistics 2002;28(4):409-445.

[5] Barzilay R, Elhadad M. Using Lexical Chains for Text Summarization. In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999; pp 111-121.

[6] Hahn U., Reimer U. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In: Inderjeet Mani and Mark T. Maybury, Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999; pp. 215-232.

[7] Humphreys BL, Lindberg DA, Schoolman HM, et al. The Unified Medical Language System: An informatics research collaboration. J Am Med Inform Assoc 1998 Jan-Feb;5(1):1-11.

[8] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001;10(Pt 1):216-20.

[9] http://www.nlm.nih.gov/medlineplus/encyclopedia.html.

[10] Rindflesch TC, Bean CA, Sneiderman CA. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc AMIA Symp 2000;:704-8.

[11] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001;:17-21

[12] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. Accepted for pub, 2003.

[13] Marcu D. From discourse structures to text summaries. Proc of the Workshop on Intelligent Scalable Text Summarization, Association for Computational Linguistics 1997:; 82-88.

[14] Jacquelinet C, Burgun A, Delamarre D, Strang N, Djabbour S, Boutin B, Le Beux P. Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation. Int J Med Inf. 2003 Jul;70(2-3):317-28.

[15] Mani I. Summarization evaluation: an overview. Proc of the 2nd meeting of the north American chapter the Association for Computational Linguistics 2001.

[16] Batagelj, A. Mrvar: Pajek - Analysis and Visualization of Large Networks. in Jünger, M., Mutzel, P., editors, Graph Drawing Software. Berlin: Springer, 2003, pp. 77-103.

Address for correspondence

Marcelo Fiszman, MD, PhD.

National Library of Medicine, Bldg 38A

8600 Rockville Pike MS 43

Bethesda, MD, USA 20894

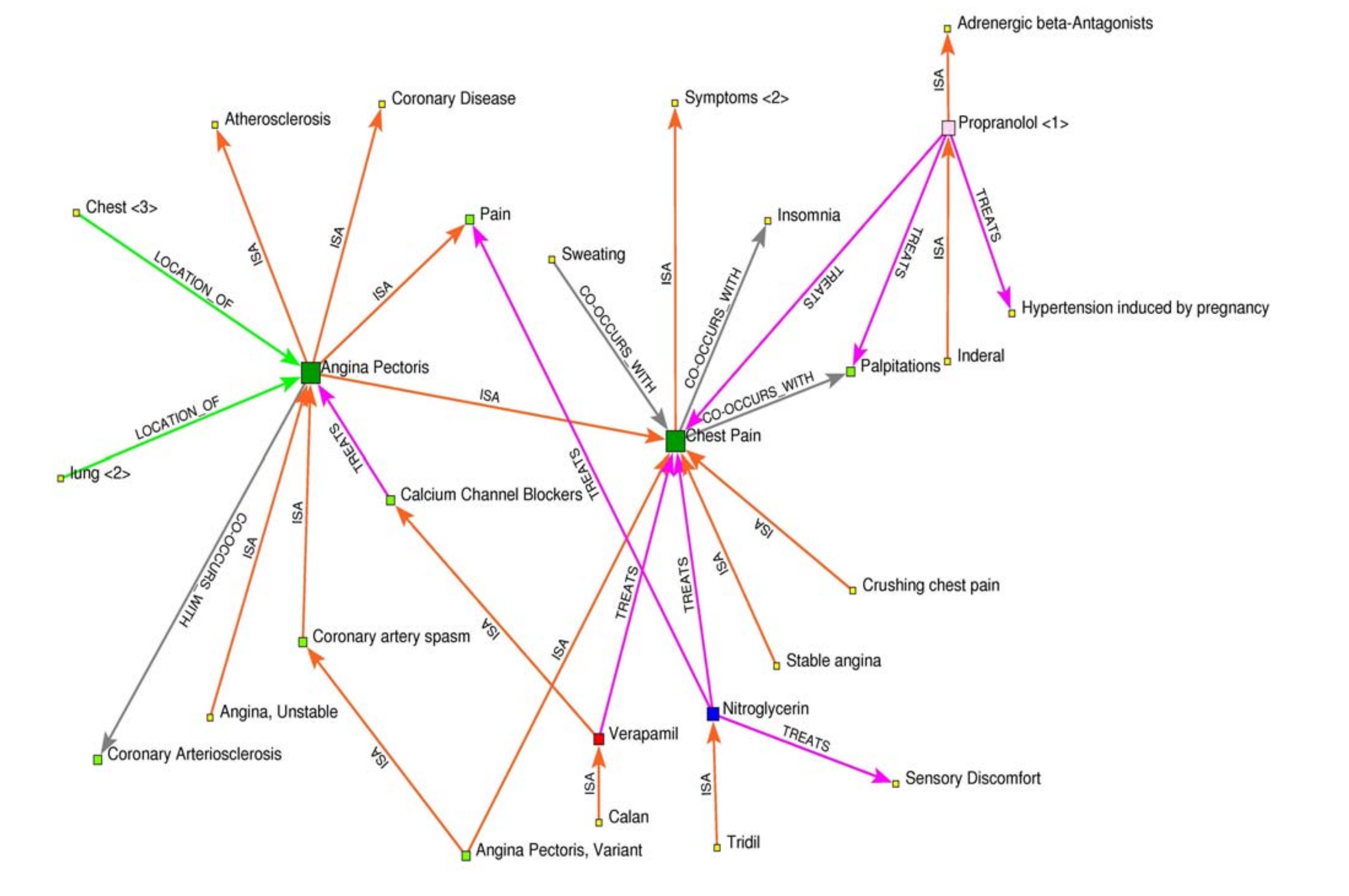

Figure 2 - Conceptual condensate for Chest Pain. Arrows follow the direction of the predicate.